

MULTI-SCALE VISUAL ATTENTION & SALIENCY MODELLING WITH DECISION THEORY

Anh Cat Le Ngo^{1,2} Li-Minn Ang³ Guoping Qiu² Kah Phooi Seng⁴

¹ Faculty of Engineering, University of Nottingham, Malaysia Campus, Malaysia

² School of Computer Science, University of Nottingham, Jubilee Campus, UK

³ Centre for Communications Engineering Research, Edith Cowan University, Australia

⁴ Department of Computer Science & Networked System, Sunway University, Malaysia

ABSTRACT

Bottom-up saliency, an early human visual processing, behaves like binary classification of interest and null hypothesis. Its discriminant power, mutual information of image features and class distribution, is closely related to saliency value by the well-known centre-surround theory. As classification accuracy very much depends on window sizes, the discriminant saliency (power) varies according to sampling scales. Discriminating power estimation in multi-scales framework needs integrating with wavelet transformation and then estimating statistical discrepancy of two consecutive scales (centre-surround windows) by Hidden Markov Tree (HMT) model. Finally, multi-scale discriminant saliency (MDIS) maps are combined by the maximum information rule to synthesize a final saliency map. All MDIS maps are evaluated with standard quantitative tools (NSS, LCC, AUC) on N. Bruce's database with ground truth data as eye-tracking locations; as well assessed qualitatively by visual examination of individual cases. For evaluating MDIS against well-known AIM saliency method, simulations are needed and described in details with several interesting conclusions, drawn for further research directions.

1. DISCRIMINANT VISUAL SALIENCY

Saliency mechanism plays a key role in perceptual organization [1]; therefore, recently several researchers attempt to generalize principles for visual saliency [2][3][4][5][6],[7]. In the decision theoretic point of view, saliency is regarded as power for distinguishing salient and non-salient classes; moreover, discriminant saliency, (DIS), combines classical centre-surround hypothesis with derived optimal saliency architecture. Saliency value at a spatial location is identified as the discriminant power of a feature set with respect to

the binary classification problem between centre and surround classes. Based on the decision theory, this approach can be generalized for variety of stimulus modalities, including intensity, color, orientation and motion [2]. Moreover, various psychophysical properties for both static and motion stimuli are shown to be accurately satisfied quantitatively by DIS saliency maps [8]. Due to ubiquity of centre-surround operator in the early stages of biological vision, bottom-up saliency is commonly defined as how certain the stimuli at each location of central visual field can be determined against other stimuli in its surround. In other words, "centre-surround" hypothesis is also a natural binary classification problem which can be solved by the well-established decision theory. In this problem, classes can be defined as follows.

- Centre class: observations within a central neighborhood W_l^1 of visual fields location l .
- Surround class: observations within a surrounding window W_l^0 of the above central region.

Feature responses are drawn from the predefined feature sets X by a random process. As there are many possible combinations and orders of how such responses are assembled, feature observations can be considered as a random process, $X(l) = (X_1(l), \dots, X_d(l))$ of dimension d . This random process is drawn conditionally on hidden variable $Y(l)$ of class states or labels (center / surround). Feature vector $x(j)$, given $j \in W_l^c, c \in \{0, 1\}$, is drawn from classes c according to the conditional probability density $P_{X(l)|Y(l)}(x|c)$ where $Y(l) = 0, 1$ are surround and centre labels. The saliency of location l , $S(l)$ is equal to the discriminant power of X for the classification of the observed feature vectors. That

discriminant concept is quantified by mutual information between feature, X and class label, Y .

$$S(l) = I_l(X; Y) = \sum_c \int p_{X,Y}(x, c) \log \frac{p_{X,Y}(x, c)}{p_X(x)p_Y(c)} dx$$

However, mutual information estimation of d -dimensional space suffers from the curse of dimensionality. Successfully tackling the problem would make information-based saliency algorithms more biologically plausible and computationally feasible. Dashan Gao and Nuno Vasconcelos have proposed a possible solution called DIS [9], which is formulated as follows.

$$I_l(X; Y) = H(Y) - H(Y|X) = \frac{1}{|W_l|} \sum_{j \in W_l} \left[H(Y) + \sum_{c=0}^1 P_{Y|X}(c|x_j) \log P_{Y|X}(c|x_j) \right] \quad (1)$$

where $H(Y) = -\sum_{c=0}^1 P_Y(c) \log P_Y(c)$ is entropy of classes Y and $-E_{Y|X} [\log P_{Y|X}(c|x)]$ is conditional entropy of Y given X . Given a location l , there are corresponding center W_l^1 and surround W_l^0 windows along with a set of associated feature responses $x(j), j \in W_l = W_l^0 \cup W_l^1$.

While DIS successfully defines discriminant saliency in information-theoretic senses, its implementation, equation 1, restrains sampled features in a single fixed-size window. Consequently, it creates a bias toward objects with distinctive features fitted in that window size. As multi-scale processing is an implicit factor of visual attention, DIS needs adapting in wavelet transform, a popular multi-resolution framework.

2. MULTISCALE FRAMEWORK

A multi-scale image binary segmentation is a great starting point for multi-scale DIS (MDIS) as it also needs to classify a data point into two classes centre, surround classes. Noted that DIS only uses the binary classification as an intermediate step to measure discriminant value. As segmentation accuracy depends on sizes of classifying windows, an appropriate choice optimizes positive classification ratio; otherwise, it leads to sub-optimal systems. For example, a large window usually provides rich statistical information and enhance reliability of the algorithm; however, it simultaneously risks including heterogeneous elements in the window, which in turn reduces segmentation accuracy. If processing with too small windows, we probably run into local maxima points while missing global meaningful points. In brief, choosing appropriate window size has vital influence on

performances of binary segmentation and consequently of DIS or MDIS.

2.1. Dyadic Classification Windows

Dynamic windows with varying sizes can be employed to obtain coarse-to-fine segmented regions [10]. Adapting this approach, MDIS can produce saliency maps with varying resolutions. In MDIS, multiscale dyadic windows are implemented due to its compact arrangement [11]; for example, an initial square image s with $2^J \times 2^J$ of $n := 2^{2J}$ pixels, the dyadic square structures can be generated by recursively dividing x into four square sub-images equally, the left-hand side of figure 1. Moreover, it is similar to the popular quad-tree structure, commonly employed in wavelet transforms, the right-hand side of figure 1. Each node of a quad-tree is a child of a node at the directly above level; meanwhile it is a parent of other nodes at the directly below level. Each node corresponds to a dyadic block, combining wavelet coefficients across different sub-bands, nodes τ in the figure 1. Let's denote each block by d_i^j given i, j are indexes of locations, levels.

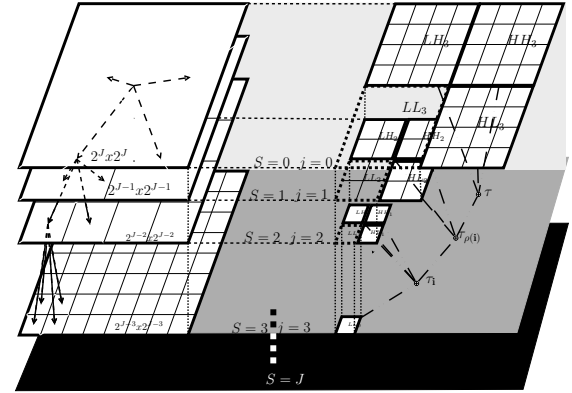


Fig. 1: Quad-tree structure

Assumed image contents are generated by random variable X , each node of the quad-tree also relates to a randomly generated block. Classification of a node into either centre or surround class requires studying its statistical property. As a node can be represented by wavelet coefficients, Gaussian Mixture Model (GMM) is utilised for estimating their likelihood from mixtures of large and small variance Gaussian distributions. Moreover, inter-scale correlation is usually found between wavelet-coefficients of different levels; hence, this statistical dependence is modelled by Hidden Markov Tree (HMT). Basically, HMT estimates likelihood of each wavelet coefficient give a hidden state, considering feature probability by GSM

and transition probability matrix. Noted that, it includes novelty and persistence elements, for which hidden states are probably changed or persisted from open scale to another. Utilization of the up-down algorithm [12] estimates likelihood $p(d_i^j|c_m)$, of all nodes given their hidden states $c_m = 0, 1$. Though binary segmentation / classification can be achieved with the maximum likelihood principle, however the results are not consistent across scales due to lack of prior information integration. Choi et al [13] proposes a Bayesian Maximum a Posterior (MAP) approach for $p(c_m|d_i^j, v_i^{j-1})$, the equation 2, whereof both parents' classes and children's features are involved in class decisions. To optimize MAP and enhance across-scale coherency, sweeping operations fuse likelihoods $f(d_i|c_i)$ along the quad-tree given the label tree prior $p(c_i^j|v_i)$.

$$\hat{c}_i^{MAP} = \operatorname{argmax}_{c_i^j \in \{0,1\}} f(c_i^j|d^j, v^j) \quad (2)$$

2.2. Multiscale Discriminant Saliency

The DIS method also uses MAP to estimate the scale parameter or variance of GGD (see section 2.4 [8] for more details) as follows.

$$\hat{\alpha}^{MAP} = \left[\frac{1}{\bar{K}} \left(\sum_{j=1}^n |x(j)|^\beta + \nu \right) \right]^{\frac{1}{\beta}} \quad (3)$$

The estimation is later included in centre / surround class decision, the equation 1. Therefore, discriminant power is strictly proportional to how difference there are between MAP values of distributions with variances α_0, α_1 from both classes. In MDIS, posterior can be computed directly by the equation 2, and its combination with mutual information principle of DIS, the equation 1 yields a multiscale estimation for discriminant power, $I_i^j(C^j; D^j)$.

$$H(C^j) + \sum_{c=0}^1 P_{C^j|D^j}(c_i^j|d^j) \log P_{C^j|D^j}(c_i^j|d^j) \quad (4)$$

Since the equation 4 yields discriminant power across scales, we can choose the maximum MAP values, $\operatorname{argmax}_j (I_i^j)$, for each location.

3. EXPERIMENTS & DISCUSSION

In our paper, we try MAP estimations with several HMT derivatives such as Universal HMT [14], Trained HMT [12], or Vector HMT [15]. Normal HMT (THMT) requires an on-line training stage

for estimating model parameters. THMT processes three wavelet orientations independently by single-variate operations; meanwhile, a vector of coefficients can be treated as multi-variate variables in similar operations by VHMT. Multi-variate nature of VHMT prefers modelling textural, especially rotation-invariant features. Though THMT or VHMT needs training stages for parameter, they could be fixed by off-line training in UHMT if general image contents are known in advance. Romberg et al. [14] have proposed a set of UHMT parameters for natural images, such approach needs evaluating against an established saliency method AIM (An Inforax Method [16]) in both quantitative (LCC, NSS, AUC, TIME [17]) or qualitative measures, visual inspection of generated saliency maps on the well-known Neil Bruce's database [18] with eye-tracking locations.

In the simulation, we deploy five dyadic scales corresponding to (U/T/V)HMT(1-5) of MDIS and integrated saliency maps are denoted by (U/T/V)HMT0. Three numerical measures linear cross correlation (LCC), normalized scan-path saliency (NSS), area under curve AUC and TIME are represented in tables 2k, 2m, 2o for (U,T,V)HMT consequently. In these tables, TIME represents computational requirement of saliency methods of (U,T,V) HMT which are listed in predictable incrementing orders. While UHMT requires the least TIME due to no requirement for training, THMT and VHMT need more computational effort for learning model parameters in single and multiple variate manners. (T,V)HMT surpass UHMT in evaluated LCC, NSS, and AUC scores, shown in the tables 2k,2m,2o and figures 2a,2f,2b. Comparatively, the proposed MDIS surpasses AIM in all quantitative measures, clearly shown by each column of these tables with **maximum** and **minimum** values. In figures 2c,2d,2e are shown the comparisons between different modes of MDIS and AIM with Receiver Operating Curve (ROC). Generally, HMT-based MDIS modes perform better than AIM in smaller scales (U,T,V)HMT(0,4,5) but MDIS in larger scales HMT(1,2,3) are equivalent or slight worse than AIM. AUC measures are increased with shrinking sizes of processing windows HMT(1-5) regardless of U/T/V modes. Meanwhile, LCC and NSS are varied more wildly, for instance, UHMT has the best LCC, NSS at the HMT4 mode; while, (T,V)HMT almost has the best evaluation at HMT0, the integrated mode. Overall, trained HMT, especially VHMT in the table 2o and figure 2j, provides more consistent numerical results through different scales. Figures 2l,2n,2p

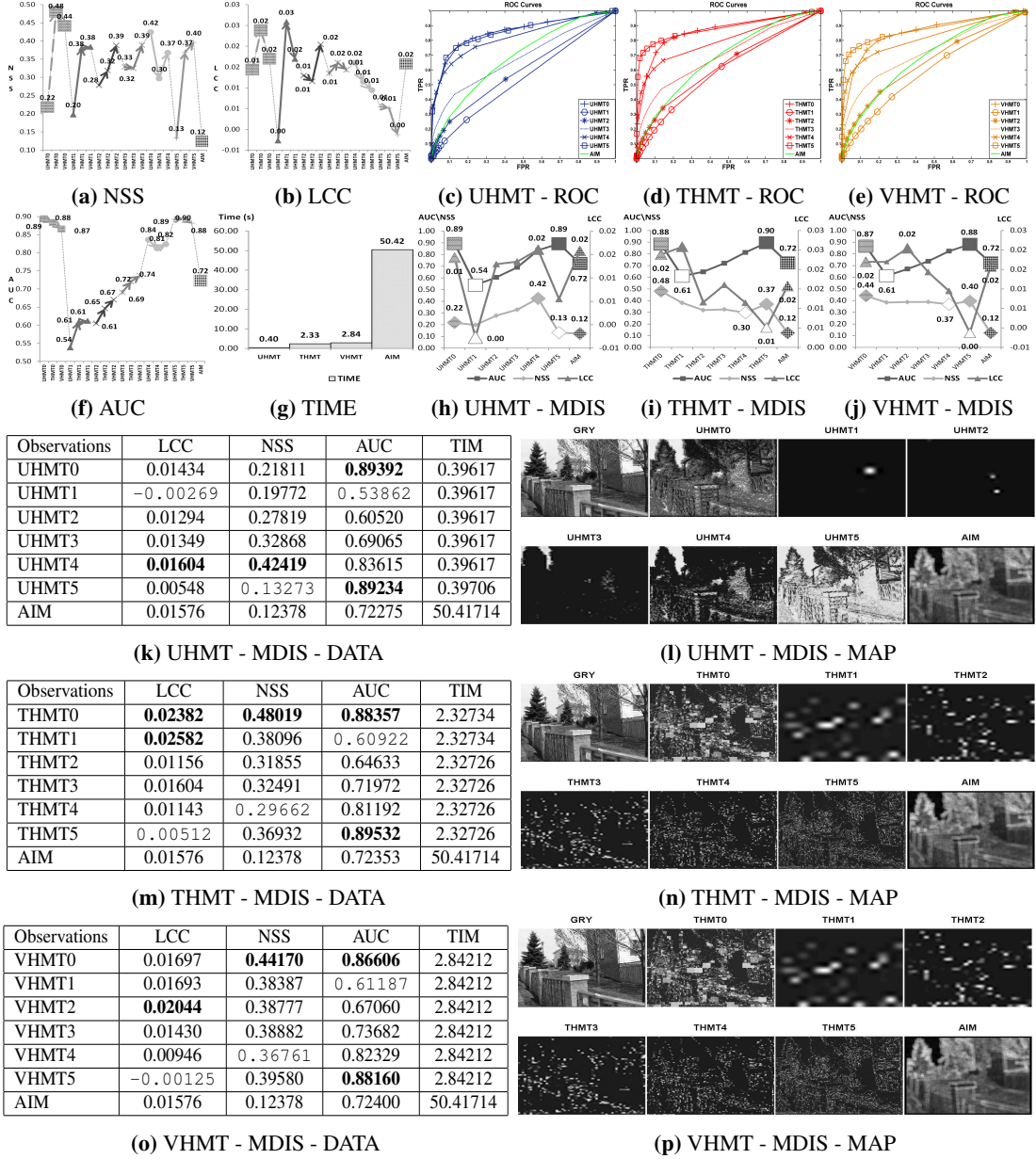


Figure 2 & Table 1: Quantitative and Qualitative evaluation of MDIS and AIM

show sample saliency maps of (U,T,V)HMT(0-5) MDISs and AIM for qualitative evaluation. (T,V)HMT have similar saliency maps while the UHMT map highlights unlikely attentive regions. Its poor performance might be due to lack of training steps.

4. CONCLUSION

In conclusion, Multiscale Discriminant Saliency (MDIS) is developed as an extension of DIS [19] under the dyadic scale framework of wavelet transform. MDIS utilizes mutual information between classes and feature distribution to quantify clas-

sifying discriminant power as saliency value in multiple dyadic-scale structures. Moreover, it fuses prior information, class decisions from previous scales, in Bayesian MAP along quad-tree in coarse-to-fine manner to create consistent saliency maps for multiple scales and final integrated map with maximum information rule. MDISs are evaluated against AIM to prove MDIS's competitiveness. For further research direction is implementation of MDIS algorithms on embedded and mobile platforms.

5. REFERENCES

- [1] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.
- [2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [3] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.
- [4] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *NIPS*, 2007.
- [5] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, p. 1–8.
- [6] Guoping Qiu, Xiaodong Gu, Zhibo Chen, Quqing Chen, and C. Wang, "An information theoretic model of spatiotemporal visual saliency," in *2007 IEEE International Conference on Multimedia and Expo*, July 2007, pp. 1806–1809.
- [7] A. C. Le Ngo, G. Qiu, G. Underwood, L. M. Ang, and K. P. Seng, "Visual saliency based on fast nonparametric multidimensional entropy estimation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, p. 1305–1308.
- [8] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, no. 1, pp. 239–271, 2009.
- [9] D. Gao and N. Vasconcelos, "Discriminant interest points are stable," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, p. 1–6.
- [10] Hyeokho Choi and R. Baraniuk, "Multiscale texture segmentation using wavelet-domain hidden markov models," in *Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems and Computers, 1998*, Nov. 1998, vol. 2, pp. 1692 – 1697 vol.2.
- [11] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532 – 540, Apr. 1983.
- [12] M.S. Crouse and R.G. Baraniuk, "Simplified wavelet-domain hidden markov models using contexts," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998*, May 1998, vol. 4, pp. 2277 –2280 vol.4.
- [13] H. Choi and R.G. Baraniuk, "Multiscale image segmentation using wavelet-domain hidden markov models," *IEEE Transactions on Image Processing*, vol. 10, no. 9, pp. 1309 –1321, Sept. 2001.
- [14] J.K. Romberg, Hyeokho Choi, and R.G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden markov models," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1056 – 1068, July 2001.
- [15] M. N. Do and M. Vetterli, "Rotation invariant texture characterization and retrieval using steerable wavelet-domain hidden markov models," *Multimedia, IEEE Transactions on*, vol. 4, no. 4, pp. 517–527, 2002.
- [16] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in neural information processing systems*, vol. 18, pp. 155, 2006.
- [17] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of Human-Model agreement in visual saliency modeling: A comparative study," *Image Processing, IEEE Transactions on*, 2012.
- [18] N. D. B. Bruce, D. P. Loach, and J. K. Tsotsos, "Visual correlates of fixation selection: a look at the spatial frequency domain," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 2007, vol. 3, p. III–289.
- [19] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," *Advances in neural information processing systems*, vol. 20, pp. 1–8, 2007.